International Journal of Computational Research and Development (IJCRD)
Impact Factor: 5.015, ISSN (Online): 2456 - 3137
(www.dvpublication.com) Volume 2, Issue 2, 2017

A COST-EFFECTIVE PRIVACY PRESERVING USING ANONYMIZATION BASED HYBRID BAT ALGORITHM WITH SIMULATED ANNEALING APPROACH FOR INTERMEDIATE DATA SETS OVER CLOUD COMPUTING

J. Sasidevi*, Dr. R. Sugumar & P. Shanmuaga Priya***

* Research Scholar, Department of Computer Science and Engineering, St. Peter's University, Chennai, Tamilnadu

** Associate Professor, Department of Computer Science and Engineering, Velammal Institute of Technology, Chennai, Tamilnadu

Cite This Article: J. Sasidevi, Dr. R. Sugumar & P. Shanmuaga Priya, "A Cost-Effective Privacy Preserving Using Anonymization Based Hybrid Bat Algorithm with Simulated Annealing Approach for Intermediate Data Sets Over Cloud Computing", International Journal of Computational Research and Development, Volume 2, Issue 2, Page Number 173-181, 2017.

Abstract:

Cloud computing is the long dreamed vision of computing as a utility, where users can remotely store their data into the cloud. In the existing research, balanced aware Fire Fly Optimization (BFFO) algorithm is used for privacy aware data set scheduling. However the existing algorithm has running time complexity and privacy preserving efficiency is reduced significantly due to lack of anonymization algorithm. To overcome the above mentioned issues, in this research, anonymization based hybrid bat algorithm with simulated annealing (AHBSA) is introduced. In the proposed system, anonymization method is used to ensure the privacy on sensitive dataset. It prevents the original information from the malicious users in the cloud environment. HBSA optimization algorithm is applied for selecting the significant attributes and achieved the privacy based on the privacy policy. It is also used to reduce the privacy preservation cost considerably. AHBSA algorithm is finding the unauthorized users and protects the intermediate dataset by providing the privacy. The efficient encryption technique is applied to provide security hence the cloud users can obtain the secret information without compromising privacy. The experimental result prove that the proposed system shows higher privacy, lower cost, lower time complexity and better anonymization by using AHBSA approach than the existing cost based Heuristic (C_HEU) and BFFO algorithms.

Key Words: Privacy Preserving, Anonymization, Intermediate Data Set, AHBSA Algorithm & Cloud Computing

1. Introduction:

Cloud computing is emerging computing model where the data owners are outsourcing their data into the cloud storage. By outsourcing the data files into the cloud, it gives many benefits to the large enterprises as well as individual users because they can dynamically increase their storage space as and when required without buying any storage devices [1]. They are: (1) the users can access the remotely stored data at anytime, from anywhere and gives permission to authorized users to share the data. (2) The users can be relieved from the burden of storage management at locally, (3) Avoidance of capital expenditure on hardware and software costs etc. To date, there are a number of cloud storage services: Amazon simple storage Space (S3), Rack space, Google, Microsoft, etc [2].

Besides, all of these advantages of outsourced data in Cloud, there are also some significant issues. One of the major issues is the privacy of outsourced data in cloud i.e., the sensitive information such as e-mail, health records, and government data may leak to unauthorized users [3].or even be hacked (Cloud Security Alliance, 2009). Since, the cloud is an open platform; it can be subjected to attacks from both malicious insiders and outsiders [4]. The Cloud service providers (CSPs) usually provide data security through mechanisms like firewalls and virtualization. However, these mechanisms do not protect users' privacy from the CSP itself due to remote cloud storage servers are untrusted.

In [5] a systematic approach, Privacy-MaxEnt, to integrate background knowledge in privacy quantification. The method is based on the maximum entropy principle. We treat all the conditional probabilities P (SA | QI) as unknown variables; it treat the background knowledge as the constraints of these variables; in addition, it also formulate constraints from the published data. The goal becomes finding a solution to those variables (the probabilities) that satisfy all these constraints. Although many solutions may exist, the most unbiased estimate of P (SA | QI) is the one that achieves the maximum entropy.

The cloud computing is emerging and developing rapidly both conceptually and in reality, the legal/contractual, economic, service quality, interoperability, security and privacy issues still pose significant challenges. In [6] describe various service and deployment models of cloud computing and identify major challenges. In particular, it discusses three critical challenges: regulatory, security and privacy issues in cloud computing. Some solutions to mitigate these challenges are also proposed along with a brief presentation on the future trends in cloud computing deployment [7] [8].

2. Related Work:

In [9] Hu et al (2017) presented a novel image service outsourcing scheme for compressive sensing (CS) reconstruction computation and identity authentication in cloud is used, which integrates the technique of CS domain processing into the secure computation outsourcing. The empirical evaluations demonstrate that the scheme has a satisfactory security and efficiency performance. Experimental results also show that identity authentication in the CS domain is feasible. Moreover, this method presents a potential scenario of identity authentication.

In [10] Li et al (2015) used a novel lightweight encryption mechanism for database (L-EncDB), which provides a secure and privacy-preserving data utilization. Such a new mechanism for database does not change the data structure after encryption and can be efficiently realize data utilization such as privacy-preserving knowledge extraction, after outsourcing database into the cloud. In this new mechanism, it utilizes a core interface provided as API to interpret SQL operations, which allows protecting sensitive information in database applications. Experimental results demonstrate that the new L-EncDB is efficient and can be applied to big database for privacy-preserving applications. Finally, it also shows L-EncDB analyze the privacy-preserving queries over encrypted NoSQL Database.

In [11] Alabdulatif et al (2017) introduces a practical framework that takes advantage of cloud resources to provide a lightweight and scalable privacy preserving anomaly detection service for sensor data. A lightweight homomorphic encryption scheme is used to ensure data security and privacy with any computational limitations overcome through a convenient data processing model that employs a single private server collaborating with a set of public servers within a cloud data centre. Virtual nodes implemented on public servers perform granular anomaly detection operations on encrypted data. The experimentation demonstrates consistently high detection accuracy with less overhead in a cloud-based anomaly detection model that is both lightweight and scalable while ensuring data privacy.

In [14] Dong et al (2014) focuses on providing a dependable and secure cloud data sharing service that allows users dynamic access to their data. In order to achieve this, it use an effective, scalable and flexible privacy preserving data policy with semantic security; by utilizing ciphertext policy attribute based encryption (CP-ABE) combined with identity-based encryption (IBE) techniques. In addition to ensuring robust data sharing security, policy succeeds in preserving the privacy of cloud users and supports efficient and secure dynamic operations including, but not limited to, file creation, user revocation and modification of user attributes. Security analysis indicates that the proposed policy is secure under the generic bilinear group model in the random oracle model and enforces fine-grained access control, full collusion resistance and backward secrecy. Furthermore, performance analysis and experimental results show that the overheads are as light as possible.

In [13] Alihodzic et al (2013) used Bat Algorithm (BA) is a biological algorithm, advanced and BA has been defined to be highly effective. BA depends on nature of echo locations for multilevel thresholds selection process which is utilizes the greater entropy criterion. The experimental output explains the BA algorithm which can search for multiple thresholds which are very nearer to the optimal ones which is defined by the exhaustive search method. The proposed system distinguished with the current algorithms, the computational times shows the performance of BA is good. This algorithm is to explain the viability of BA method for multilevel threshold. Also, it gives the new option to the traditional methods because of its simplicity and efficiency.

In [14] Emary et al (2014) used a hybrid Bat with rough set feature selection method to select a smaller number of features and achieving similar or even better classification performance than using all features. Different initialization methods namely normal, large, small, and mixed initialization methods are used to initialize the different optimizers. BA is used for searching the feature space for minimal feature size and maximum classification performance. BA proves performance advance in both classification accuracy and feature reduction over common methods such as Particle Swarm Optimization (PSO) and genetic algorithm. Also, BA proves its capability to converge to a good-enough solution. The obtained result proves the robustness of BA regardless of the used initialization method. The used fitness function combines both classification quality and reduction in feature set size and hence targets both.

In [15] Deviet al (2015) presented a simulated annealing procedure is used to find the feature subset for each class so that patterns of this class are classified correctly. To classify a test pattern, combinations of k classifiers are used one for each class. While more bookkeeping is required as it is necessary to store the feature subset for each class, the feature selection is done only once and stored. The use of class-based feature selection is found to give good classification accuracy. In such a case only a few features maybe important for a particular class. In addition, the features selected for each class are an indication of which features are important for a class. It gives a description of a class. Class-based feature selection needs to be tried out with more data sets in the future, especially datasets where the number of features and the number of classes are large. It is also possible to use other classifiers besides the nearest neighbor bases approaches to carry out class specific feature selection.

3. Proposed Methodology:

In the proposed system, the HBSA algorithm is used to increase the privacy preserving and reduce the cost function on cloud environment. The overall proposed system is illustrated in the fig 1.

3.1 Problem Analysis: Data provenance is employed to manage intermediate data sets in this research. Provenance is commonly defined as the origin, source or history of derivation of some objects and data, which can be reckoned as the information upon how data were generated [15].Reproducibility of data provenance can help to regenerate a data set from its nearest existing predecessor data sets rather than from scratch. It assumes herein that the information recorded in data provenance is leveraged to build up the generation relationships of data sets.

Threat Model: In threat model, It considers mainly two types of threats, which are disturbing the outsourced data in the cloud: Internal Attacks and External Attacks 1.InternalAttacks: which are initiated by malicious insiders: Cloud users, malicious third party user (either cloud provider or customer organizations) are self-interested to accesses the data or disclose the data stored in the cloud. They also alter or modify the data.2. **External Attacks:** which are initiated by unauthorized outsiders, it assumes that external attackers can compromise all storage servers, so that they can intentionally access the owner's data. It defines several basic notations below. Let do be a privacy-sensitive original data set. It uses D <d1;d2; . . . ; dng to denote a group of intermediate data sets generated from do where n is the number of intermediate data sets. Note that the notion of intermediate data herein refers to both intermediate and resultant data. Directed Acyclic Graph (DAG) is exploited to capture the topological structure of generation relationships among these data sets. A DAG representing the generation relationships of intermediate data sets D from do is defined as a Sensitive Intermediate dataset Graph, denoted as SIG.

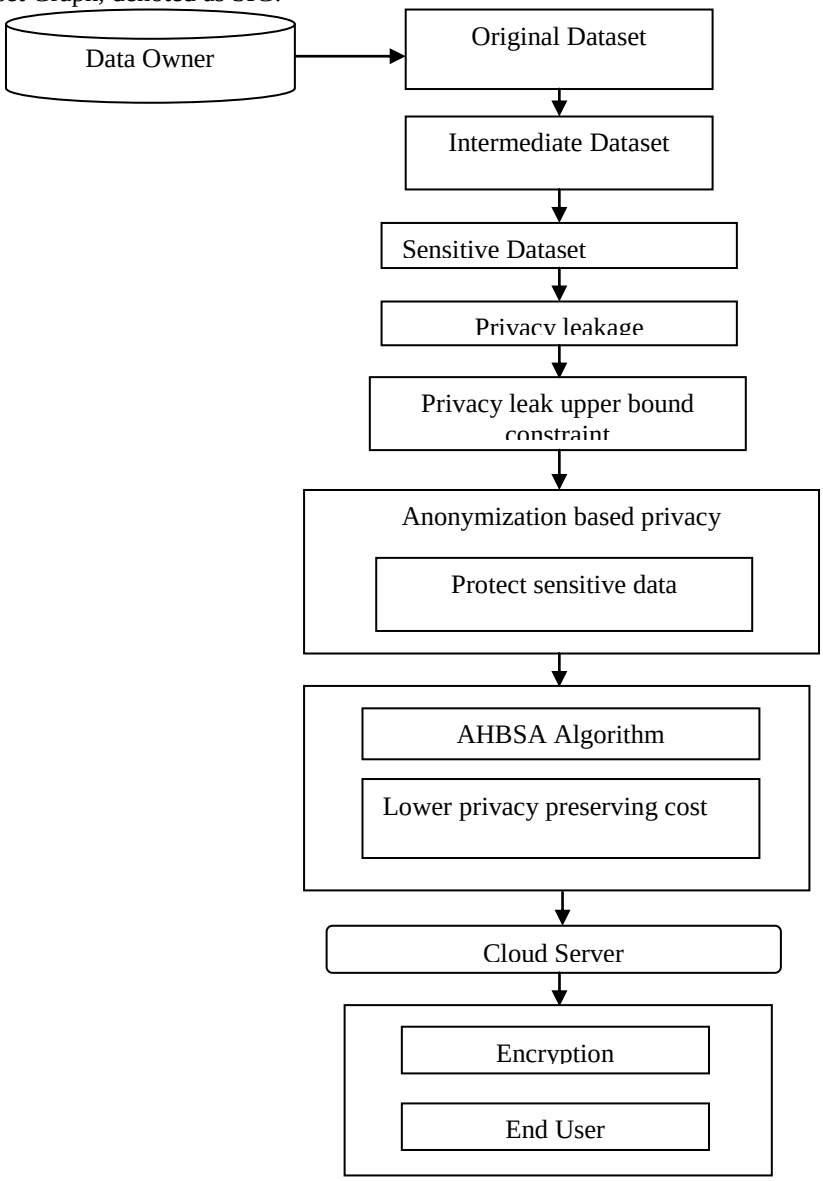

Figure 1: Overall block diagram of the proposed system

3.2 Privacy Preservation for Intermediate Data Set:

The value of the joint privacy leakage incurred by multiple data sets in $D = \{d_1, d_2, \dots, d_n\}, n \in N$ is defined by

$$PL_m(D) \triangleq H(S,Q) - H_D(S,Q) \quad (1)$$

$H(S,Q)$ and $H_D(S,Q)$ are the entropy of $< S,Q)$ before and after data sets in D are observed, respectively. $H(S,Q) = \log(|QI|) . |SD|). H_D(S,Q)$ can be calculated once $P(S,Q)$ is estimated after data sets in D are observed. Given the relationship between ε and $PL_m(D^{une})$ in PLC, ε rangesin the interval

$$[max_{1 \leq i \leq n}\{PL_s(d_i)\}, \log(|QI|.|SA|)]$$

It is focused to derive an upper bound of $PL_m(D^{une})$ thatcan be easily computed. Intuitively, if an upper bound $BPL_m(D^{une})$ is found, a stronger privacy leakage constraint $B(PL_m(D^{une})) \leq \varepsilon$ can be a sufficient condition of thePLC. Accordingly, $PL_m(D^{une})$ will never exceed the threshold ε if $B(PL_m(D^{une})) \leq \varepsilon$ holds.

Let d_u and d_v be two data sets whose privacy leakage are $PL_s(d_u)$ and $PL_s(d_v)$ respectively. The joint privacy leakage caused by them together is $PL_m(d_u, d_v)$. This property of joint privacy leakage can be extended to multiple unencrypted data sets in D^{une}

$$PL_m(D^{une}) \leq \sum_{d_i \in D^{une}} PL_s(d_i)$$

An upper bound constraint-based approach to select the necessary subset of intermediate data sets that needs to be encrypted for minimizing privacy-preserving cost. To satisfy the PLC, we decompose the PLC recursively into different layers in an SIT. Then, the problem stated in (3) can be addressed via tackling a series of small-scale optimization problems. Let the privacy leakage threshold required in the layer L_i be ε_i, $1 \leq i \leq H$.The privacyleakage incurred by UD_i in the solution π_i can never be larger than ε_i, i.e., $PL_m(UD_i) \leq \varepsilon_i$. The threshold ε_i can beregarded as the privacy leakage threshold of the remainderpart of an SIT after the layer L_{i-1}. In terms of the basic idea of approach, the privacy leakage constraint $PL_m(UD_i) \leq \varepsilon_i$ is substituted by one of its sufficientconditions.

According to (7), the PLC can be substituted by a set of privacy leakage constraints, named as PLC1:

$$\sum_{d \in UD_i} PL_s(d) \leq \varepsilon_i, 1 \leq i \leq H \quad (2)$$

The above threshold $\varepsilon_i, 1 \leq i \leq H$ is calculated by

$$\begin{cases} \varepsilon_i = \varepsilon_{i-1} - \sum_{d \in UD_i} PL_s(d) \\ \varepsilon_1 = \varepsilon \end{cases} \quad (3)$$

A local encryption solution in the layer Li is feasible if it satisfies the PLC_1. The set of feasible solutions in L_i isdenoted as $\Lambda_i^f \triangleq \{\pi_{ij} | \pi_{ij} \epsilon \Lambda_i$, where j is the number of feasible solutions. Similarly, a feasible global encryption solution can be denoted as $\pi_f^k \triangleq < \pi_{1_{j1}}, \dots \dots \pi_{H_{jH}} >$, where $\pi_{ij} \in \Lambda_i^f, 1 \leq i \leq H, 1 \leq k \leq \prod_{i=1}^{H} |\Lambda_i|$ Given a feasible global solution π_f^k for an SIT, compress the SIT into a "compressed" tree layer by layer from L_1 to L_H denoted as $CT(\pi_f^k)$, where H is the height of the DG in SIT. The construction of $CT(\pi_f^k)$ is achieved via three steps.

First, the data sets in EDi are "compressed" into one encrypted node. According to the EDT property, these compressed nodes together with the original data set appear to be a string with the length being H. Second, all offspring data sets of the data sets in UDi are omitted. This will not affect the privacy preserving in terms of the RPC property. Third, the data sets in UDi are compressed into one node.

Usually, more than one feasible global encryption solution exists under the PLC1 constraints, because there are many alternative local solutions in each layer. Further, each intermediate data set has various size and frequency of usage, leading to different overall cost with different solutions. Therefore, it is desired to find a feasible solution with the minimum privacy-preserving cost under privacy leakage constraints. Note that the minimum solution mentioned herein is somewhat pseudo minimum because an upper bound of joint privacy leakage is just an approximation of its exact value. But a solution can be exactly minimal in the sense of the PLC1 constraints. It derives the recursive minimal cost formula as follows.

The minimum cost for privacy preserving of the data sets after L_{i-1} under the privacy leakage threshold ε_i isrepresented as $CM_i(\varepsilon_i)$, $1 \leq i \leq H$.Given a feasible localencryption solution $\pi_i =< ED_i, UD_i >$ in L_i the cost incurredby the encrypted data sets in Li is denoted as $C_i(\pi_i)$

$$C_i(\pi_i) \triangleq \sum_{d_k \in ED_i} S_K . PR. f_k, \quad 1 \leq i \leq H \quad (4)$$

Then $CM_i(\varepsilon_i)$ is calculated by the recursive formula

$$\begin{cases} CM_i = \min_{\pi_{ij} \in \Lambda_f^k} (\sum_{d_k \in ED_i} S_k . PR. f_k \quad) \\ \quad + CM_{i+1}(\varepsilon_i - \sum_{d_k \in ED_i} PL_s(d_k)) \} \\ \quad CM_{H+1}(\varepsilon_{H+1}) = 0 \end{cases} \quad (5)$$

3.3 Anonymization Method for Privacy:

Anonymization is done to remove the personally identifiable information from data sets to preserve privacy. Anonymization of data preserves the original structure and field layout of the data thus the data looks original and realistic.

Data anonymization approach is based on agglomerative hierarchical clustering algorithm in this research. Agglomerative hierarchical clustering is a bottom-up clustering method. It starts with every single tuple in a single cluster. Then, in each successive iteration, it merges the closest pair of clusters by satisfying some similarity criteria until some stopping rule is satisfied. The critical issue for an agglomerative hierarchical clustering algorithm is to choose the optimal pair of clusters for merging among a large amount of clusters in each iteration. In this method for data anonymization, it make the decision by both of the two factors including information loss and impurity gain. The pair of clusters is chosen for merging, if the merging causes minimum information loss and maximum impurity gain. So, merging index, denoted by MI, is defined in our method to measure the quality of a pair of clusters on both features of information loss and impurity gain.

- ✓ Given a dataset *D*, parameters *k* and *l*, build initial cluster-set according to the values of QI. For clusters in initial cluster-set, assign those meet k-l constraint to G1, and others are assigned to *G*0, where *G*1 and *G*0 are cluster sets;
- ✓ For the clusters in G0 and G1, compute MI for each pair of clusters (Ciand Cj, and at least one of them is from G0, namely, doesn't meet k-l constraint). Choose the pair of clusters with the maximum of MI, and merge them to a new cluster *C*;
- ✓ Compute the centroid of C, and update each tuple in C with replacing the value of QI by the centroid of C. Assign C to G1 if it meets k-l constraint, otherwise, assign it to G0;
- ✓ Repeat steps 2 and 3, until G0= Ø. Suppose that an original dataset *D* has been anonymized to *D'*, and tuple $t_i^{'}$in $D^{'}$corresponds to $t_i$in *D'*, namely t_i is anonymized to $t_i^{'}$. The information loss of the map from$t_i$to $t_i^{'}$,denoted by $IL(t_i)$, is defined as follows:

$$IL(t_i) = \sqrt{\sum_{j=1}^{v}(t_{ij} - t_{ij}^{'})^2}$$

Where *v* is the number of attributes,t_{ij} is the value of the j[th] attribute of$t_i^{'}$. Obviously, $IL(t_i)$ is between 0 and *v*. Thus, the utility loss, denoted by *UL*, can be defined as the average information loss of dataset *D*:

$$UL = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{v}\left(\sqrt{\sum_{j=1}^{v}(t_{ij} - t_{ij}^{'})^2}\right)$$

Where *n* is the number of tuples in *D*. *UL* is a value between 0and 1.

The quantification used to measure data privacy is the degree of uncertainty, according to which original private data can be inferred. For various types of anonymization algorithms, the degree of privacy is estimated in different ways. It uses a universal measure of data privacy based on the concept of information entropy. The change on information entropy is used to measure the privacy level. Domingo-ferrer addresses that the privacy level can be assessed by using the disclosure risk, that is, the risk that a piece of information be linked to a specific tuple. This method measures the privacy level by the amount of extra information the adversary learn from the anonymized data.

3.5 HBSA Optimization Algorithm Based Feature Selection:

In this section, EB algorithm is introduced for efficient model training. After extraction of the feature vectors it will be applying the classification algorithms to recognize the features from the intermediate dataset. The emotion recognition rate is dependent on the types of features extracted and the selection of the classification algorithm. In this research, it has been evaluated that the HBSA algorithms are better optimization algorithm.

Behaviour of Microbats: Most of microbats have advanced capability of echolocation. These bats can emit a very loud and short sound pulse; the echo that reflects back from the surrounding objects is received by their extraordinary big auricle. Then, this feedback information of echo is analyzed in their subtle brain. They not only can discriminate direction for their own flight pathway according to the echo, but also can distinguish different insects and obstacles to hunt prey and avoid a collision effectively in the day or night.

All bats use echolocation to sense distance, and they also "know" the difference between food/prey and background barriers in some magical way [17].

Bats fly randomly with velocity Vi at position xi with a fixed frequency $f$min, varying wavelength λ, and loudness *A*0 to search for prey. They can automatically adjust the wavelength (or frequency) of their emitted pulses and adjust the rate of pulse emission $r\in$ [0, 1], depending on the proximity of their target.

Although the loudness can vary in many ways, it assumes that the loudness varies from a large (positive) *A*0 to a minimum constant value *A*min.

In addition, for simplicity, they also use the following approximations: in general, the frequency f in a range [$f$min, $f$max] corresponds to a range of wavelengths [λmin, λmax]. In fact, they just vary in the frequency while fixed in the wavelength λ and assume $f\in$ [0, $f$max] in their implementation. This is because λ and f are related due to the fact that λf = V is constant.

In simulations, they use virtual bats naturally to define the updated rules of their positions xi and velocities Vi in a D-dimensional search space. The new solutions xti and velocities Vti at time step t are given by

$$f_i = f_{min} + (f_{max} - f_{min})\beta \quad (6)$$
$$v_i^t = v_i^{t-1} + (x_i^t - x_*)f_i \quad (7)$$
$$x_i^t = x_i^{t-1} + v_i^t$$

Where $\beta \in [0, 1]$ is a random vector drawn from a uniform distribution. Here, $x*$ is the current global best location (solution) which is located after comparing all the solutions among all the n bats.
For the local search part, once a solution is selected among the current best solutions, a new solution for each bat is generated locally using random walk:

$$x_{new} = x_{old} + \varepsilon A_t \quad (8)$$

Where $\varepsilon \in [-1, 1]$ is a random number, while $At = \langle Ati \rangle$ is the average loudness of all the bats at this time step.
The loudness Ai and the rate ri of pulse emission have to be updated accordingly as the iterations proceed. These formulas are

$$A_i^{t+1} = \alpha A_i^t$$
$$r_i^{t+1} = r_1^0[1 - \exp(-\gamma t) \quad (9)$$

Where α and γ are constants
Levy Flights: Levy flights are Markov processes, which ´ differ from regular Brownian motion, whose individual jumps have lengths that are distributed with the probability density function (PDF) $\lambda(x)$ decaying at large x as $\lambda(x) = |x|-1-\alpha$ with $0<\alpha$

- ✓ Stability: distribution of the sum of independent identically distributed stable random variables equal to distribution of each variable.
- ✓ Power law asymptotics ("heavy tails").
- ✓ Generalized Central Limit Theorem: The central limit theorem states that the sum of a number of independent and identically distributed (i.i.d.) random variables with finite variances will tend to a normal distribution as the number of variables grows.
- ✓ Which has an infinite variance with an infinite mean values

Algorithm:
Objective function $f(x) = [x_1, x_2, \ldots . x_d]^T$
Initialize the bat population $x_i = (i = 1,2, \ldots . , n)$ and v_i
Define pulse frequency f_i and x_i initialize pulse rates r_i and the loudness A_i
While (t<Max number of iterations)
For each Bat b_i do
Generate new features by adjusting frequency, and updating velocities (descriptors) and locations/solutions (8)
if (rand>r_i)
Select attributes among the best feature solutions
Generate a local solution around the selected best solution
End if
Generate a new solution by flying randomly
Apply simulated annealing
Randomly select a neighbor feature from the given input
If(rand<A_i&$f(x_i) < f(x_*)$)
Accept the new features remove the noise features
Increase r_i and reduce A_i
End if
Rank the bats and find the current best x_i
End while
Post process results and visualization
The EB algorithm used to select the best attributes from the given dataset. The optimal features are generated by generating the best fitness values. The simulated annealing approach needs an initial solution as well. It randomly selected a feasible solution and used it as an initial solution. In the meantime, the neighboring solutions for a given solution are defined as binary vectors with one bit different from the given solution. Thus it is used to increase the privacy preservation considerably.

3.6 Ensure Privacy Using Efficient Scheme:

The privacy is concerned with the difficulty level of gaining information about sensitive attributes for an adversary. The higher the degree of uncertainty achieved by anonymization algorithm, the more difficulties the adversary faced, and the better the data privacy is protected. And the uncertainty of a dataset can perfectly be measured by impurity. So, in this section, it uses the metric impurity to assess the privacy level of anonymized data.

Given a dataset D with n tuples distributed in h equivalence classes, $C1, C2, \ldots, Ch$. And for each equivalence $Ci(i = 1,2, \ldots, h)$, it consists of tuples and its impurity is impurity (Ci). Then the summary impurity of D, denoted by impurity (D), is defined as follows:

$$impurity\ (D) = \sum_{i=1}^{h} \frac{n_i}{n} impurity(C_i)$$

D is an initial dataset with the lowest privacy. In this case, *D* consists of a large amount of equivalence classes because there are a large number of distinct values on *QI*. And most of the equivalence classes consist of only one tuple, which means the impurity of each class, impurity (Ci), equals 0. So, the summary impurity of D is approximately equal to 0.

D is a completely anonymized dataset with the highest privacy. A completely anonymized dataset is that all the tuples in the dataset are generalized to be with the same value on *QI*. In this case, *D* consists of only one equivalence with the largest summary impurity denoted by impurity max (*D*). Anyway, a completely anonymized dataset is useless for the largest information loss.

The given an initial dataset *D* which has been anonymized *D'*, the privacy gain of anonymization, denoted by *PG*, is defined on the basis of impurity max (D) as follows:

$$PG = \frac{impurity\ (D')}{impurity_{max}\ (D)}$$

PG is between 0 and 1, which represents the privacy level we achieve in *D'*, compared to $impurity_{max}\ (D)$.

4. Experimental Result:

U-Cloud is a cloud computing environment at the University of Technology Sydney (UTS). The system overview of U-Cloud is depicted in Fig. 4. The computing facilities of this system are located among several labs at UTS. On top of hardware and Linux operating system, it installs KVM virtualization software which virtualizes the infrastructure and provides unified computing and storage resources. To create virtualized data centres, install Open Stack open-source cloud environment for global management, resource scheduling and interaction with users. Further, Hadoop is installed based on the cloud built via Open Stack to facilitate massive data processing. The experiments are conducted in this cloud environment.

Original dataset

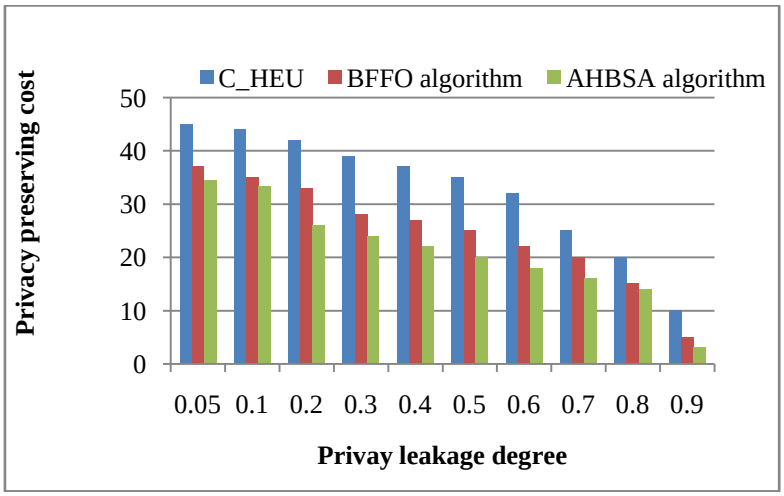

Figure 2: Privacy preserving cost vs privacy leakage degree

ε_d =0.01

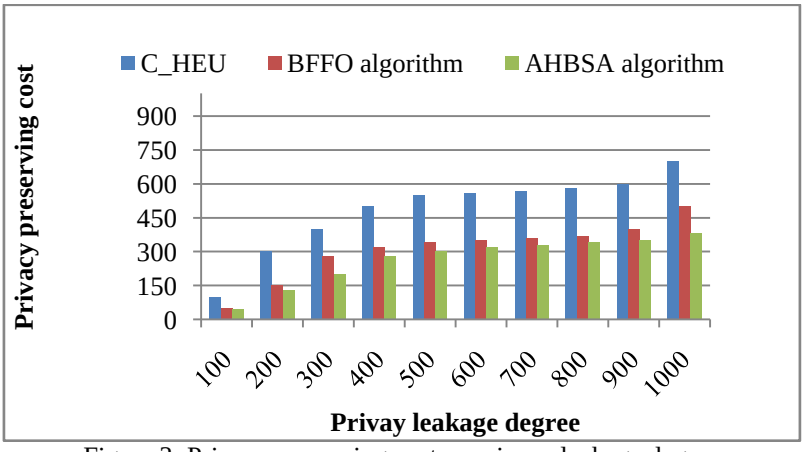

Figure 3: Privacy preserving cost vs privacy leakage degree

ε_d =0.05

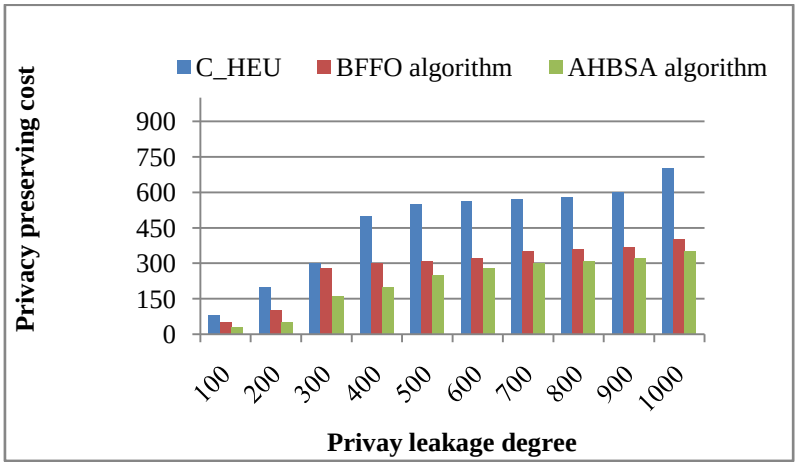

Figure 4: Privacy preserving cost vs privacy leakage degree

ε_d =0.1

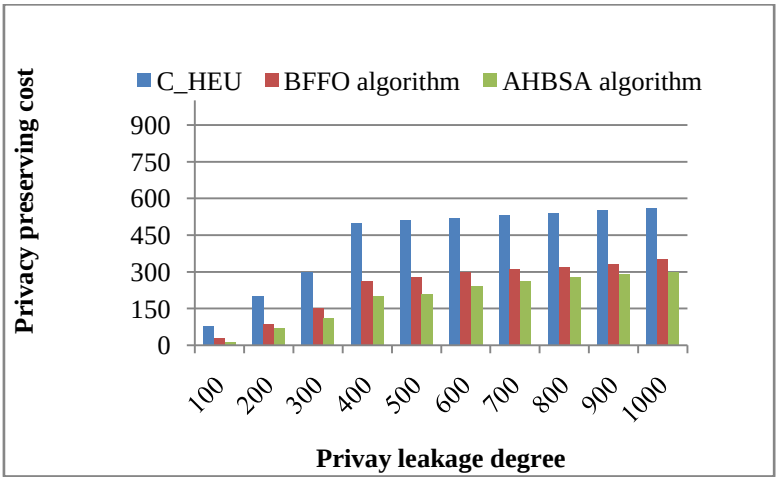

Figure 5: Privacy preserving cost vs privacy leakage degree

ε_d =0.2

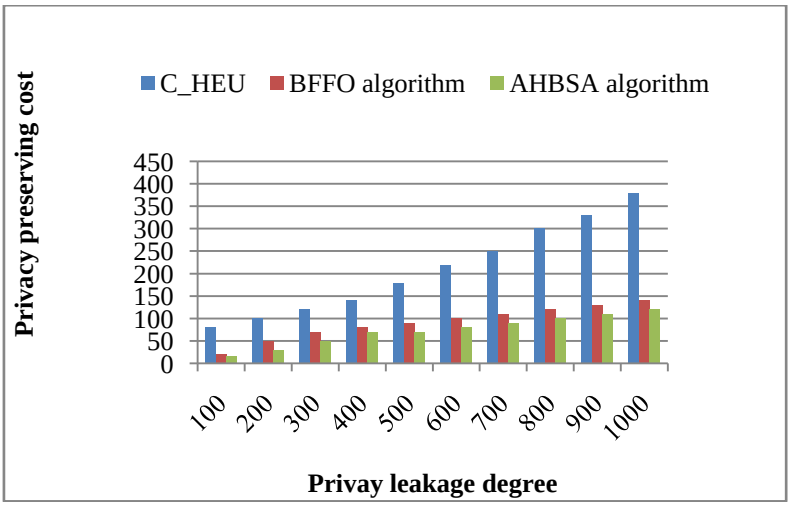

Fig 6 Privacy preserving cost vs privacy leakage degree

From the fig 2, the proposed AHBSA algorithm shows the much lower privacy preserving cost than the different privacy leakage degree. From the fig 3, different privacy leakage degree and privacy cost metric is computed for $\varepsilon_d = 0.01$. When the number of intermediate data sets is getting larger, more datasets are required to be encrypted. The privacy cost is reduced significantly using the BFFO algorithm. The fig 4, 5 and 6 are plotted $\varepsilon_d = 0.05$,$\varepsilon_d = 0.1$ and $\varepsilon_d = 0.2$ respectively which shows the lower privacy cost for various privacy leakage degree considerably. Thus the result concludes that the proposed system is better in terms of lower cost and better privacy.

5. Conclusion:

Cloud computing is the technology which enables obtaining resources like so services, software, hardware over the internet. This proposed method is used to protect the intermediate dataset privacy by using the anonymization based hybrid optimization algorithm more effectively. In this work, use a new data anonymization approach based on impurity gain and hierarchical clustering to resist probabilistic inference attacks. Anonymization method is applied to protect the sensitive information efficiently. Then, apply the HBSA optimization algorithm to reduce the cost and time complexity significantly on the intermediate datasets. This research is used to evaluate the quality of anonymization results, which can help a publisher to choose an optical tradeoff between utility and privacy. The cloud data are encrypted and packaged with a usage policy. The data when accessed will consult its policy, create a virtualization environment, and attempt to assess the trustworthiness of the data environment. The experimental result proves that the proposed AHBSA optimization approach has better performance rather than the existing BFFO and C_HEU system in terms of lower privacy cost, lower time complexity and better privacy.

6. References:

1. Pasupuleti, Syam Kumar, Subramanian Ramalingam, and Rajkumar Buyya. "An efficient and secure privacy-preserving approach for outsourced data of resource constrained mobile devices in cloud computing." Journal of Network and Computer Applications 64 (2016): 12-22.
2. Jaeger PT, Lin J, Grimes JM. Cloud computing and information policy: computing in a policy cloud? J Inform TechnolPolit 2009;5(3):269–83.
3. Slocum Z. Your Google docs: soon in search results?; 2009. ⟨http://news.cnet.com/ 8301-17939_109-10357137-2.html⟩.
4. Haclgiimfi H, Iyer B, Li C, Mehrotra S. Executing SQL over encrypted data in database-service-provider model, Technical Report TR-DB0202. Irvine: Database Research Group at University of California; 2002.
5. Du, Wenliang, ZhouxuanTeng, and Zutao Zhu. "Privacy-maxent: integrating background knowledge in privacy quantification." Proceedings of the 2008 ACM SIGMOD international conference on Management of data. ACM, 2008.
6. Sen, Jaydip. "Security and privacy issues in cloud computing." Architectures and Protocols for Secure Information Technology Infrastructures (2013): 1-45.
7. Kumar, K. Kiran, K. Padmaja, and P. Radha Krishna. "Automatic protocol blocker for privacy-preserving public auditing in cloud computing."International Journal of Computer science and Technology 3 (2012): 936-940.
8. Ruj, Sushmita, Milos Stojmenovic, and AmiyaNayak. "Privacy preserving access control with authentication for securing data in clouds." Cluster, Cloud and Grid Computing (CCGrid), 2012 12th IEEE/ACM International Symposium on. IEEE, 2012.
9. Hu, Guiqiang, et al. "A Compressive Sensing based privacy preserving outsourcing of image storage and identity authentication service in cloud."Information Sciences 387 (2017): 132-145.
10. Li, Jin, et al. "L-EncDB: A lightweight framework for privacy-preserving data queries in cloud computing." Knowledge-Based Systems 79 (2015): 18-26.
11. Alabdulatif, Abdulatif, et al. "Privacy-preserving anomaly detection in cloud with lightweight homomorphic encryption." Journal of Computer and System Sciences (2017).
12. Dong, Xin, et al. "Achieving an effective, scalable and privacy-preserving data sharing service in cloud computing." Computers & security 42 (2014): 151-164.
13. Alihodzic, Adis, and Milan Tuba. "Bat algorithm (BA) for image thresholding."Recent Researches in Telecommunications, Informatics, Electronics and Signal Processing (2013): 17-19.
14. Emary, E., WaleedYamany, and Aboul Ella Hassanien. "New approach for feature selection based on rough set and bat algorithm." Computer Engineering & Systems (ICCES), 2014 9th International Conference on. IEEE, 2014.
15. Devi, V. Susheela. "Class Specific Feature Selection Using Simulated Annealing." International Conference on Mining Intelligence and Knowledge Exploration. Springer International Publishing, 2015.